Data Deduplication Methods: A Review Paper

[1]Vijayata Ramteke, [2]Naziya Pathan, [3]Prachi Bhure
[1]M.Tech Student, [2]Assistant Professor, [3]Assistant Professor
[1]Computer Science & Engineering,
[1]Nuva College of Engineering & Technology, Nagpur, India

Abstract: The cloud storage services are used to store intermediate and persistent data generated from various resources including servers based networks. The outcome of such developments is that the data gets duplicated and gets replicated rapidly especially when large numbers of cloud users are working in a collaborative environment to solve large scale problems in geo-distributed networks. The data gets prone to breach of privacy and high incidence of duplication. When the dynamics of cloud services change over period of time, the ownership and proof of identity operations also need to change and work dynamically for high degree of security. In this work we will study the following concepts, methods and the schemes that can make the cloud services secure and reduce the incidence of data duplication. With the help of cryptography mathematics and to increase potential storage capacity. The proposed scheme works for deduplication of data with arithmetic key validity operations that reduce the overhead and increase the complexity of the keys so that it is hard to break the keys.

IndexTerms **- Deduplication, Arithmetic Validity, Proof of Ownership, Key Management, Zero Proof Algorithm.**

I. INTRODUCTION

Online storage providers require sophisticated algorithms for managing their data centers storage. One of the methods employed to increase the storage efficiency is “Deduplication”. But multiple challenges impact the overall operational difficulty in running such data center [1] operations. The foremost challenge is a security of the individual file while undergoing deduplication process. Most of these organizations need to emphasize on technologies related to encryption [2] (SSL, AES etc. and secure user password interaction). The user management consist of many component which include key management and key validation process. Without their operations deduplication would remain unsecure process and file would always remain under multiple thread including integrity loss and breach of privacy.

In a distributed database management systems special
care is taken to avoid duplication of data either by minimizing the number of writes for saving I/O bandwidth or de normalization. Databases use the concept of locking to avoid ownership issues, access conflicts and duplication issues. But even as disk storage capacities continue to increase and are becoming cheaper, the demand for online storage has also increased many folds. Hence, the cloud service providers (CSP) continue to seek methods to reduce cost of Deduplication and increase the potential capacity of the disk with better data management techniques. The data managers may use either compression or deduplication methods to achieve this business goal. In broad terms these technologies can be classified as data reduction techniques.

II. BACKGROUND

This section gives the description of methods employed for reducing duplicate data payload and issues related to ownership, trust and deduplication mechanism in cloud services [3].

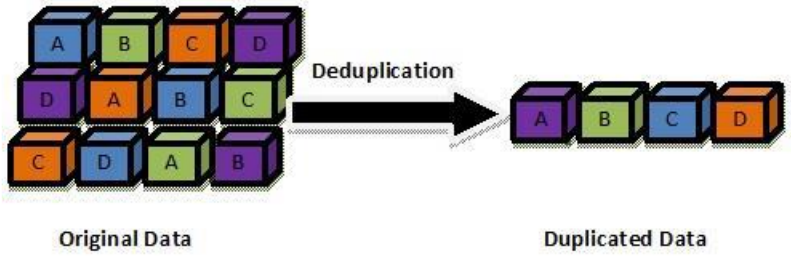

Fig.1. Deduplication Process

A. Data Reduction

The purpose is to obtain a reduced representation of a data set file that much smaller in volume, yet provide a same configuration even, if the data is modified in a collaborative environment. The reduced representation does not necessarily means a reduction in size of the data, but reduction in unwanted data or duplicates [4] is the existence of the data entities. In simple words the data reduction process would retain only one copy of the data and keep pointers to the unique copy if duplicates are found. Hence data storage [5] is reduced.

B. Compression [4]

It is a useful data reduction method as it helps to reduce the overall resources required to store and transmit data over network medium. However, computational resources are required for data reduction method. Such overhead can easily be offset due to the benefit it offers due to compression. However, an subject to the space time complexity trade off, for example, a video compression may require expensive investment in hardware, for its compression and decompression and viewing cycle, but it may help to reduce space requirements in case there is need to achieve the video.

C. Deduplication [3]

Deduplication is a process which typically consist of steps that divide the data into data sets of smaller chunk sizes and use an algorithm to allocate each data block a unique hash code. In this, the deduplication process further find similarities between the previously stored hash codes to determine if the data block is already in the storage medium. Few methods use the concept comparing back up to the previous data chunks at bit level for removing obsolete data.

 Source Based Deduplication: It is a case when deduplication is initiated at the original location where the data resides.

 Target Based Deduplication: Target based deduplication is a case when data is first transported to a target disk or storage location and then deduplication process starts. It requires higher bandwidth and uses virtual table libraries or intelligent disk transfer systems to complete the process.

 Global and Local Based Deduplication: Global deduplication functions when a single system can process large number of files across the entire enterprise rather than across each system. Global deduplication systems may provide better reduction ratios but they are normally lower than standalone (local) deduplication system.

 Semantic Based Deduplication: It is a multi- layered approach of deduplication, where level of deduplication is configurable and data can be processed in multiple stages globally. The data processing is done on the basis of Master – slave and in the case there is error in execution of operation. The master –slave mechanism is configure to check the operation accordingly.

 Software and Hardware Based Deduplication: When the focus of backup and elimination of redundant file, methods, uses and high grade hardware rather than virtualization, software
defined processes or parallelization of tasks is called hardware based deduplication.

 Hybrid Based Deduplication: When the deduplication system takes the advantage of high grade hardware as well as software capabilities it is called hybrid approach. It is usually adopted when there is a need for processing huge data bases and the enormous storage needs to be done in real time. Such solutions are enterprise solutions and are expensive in nature.

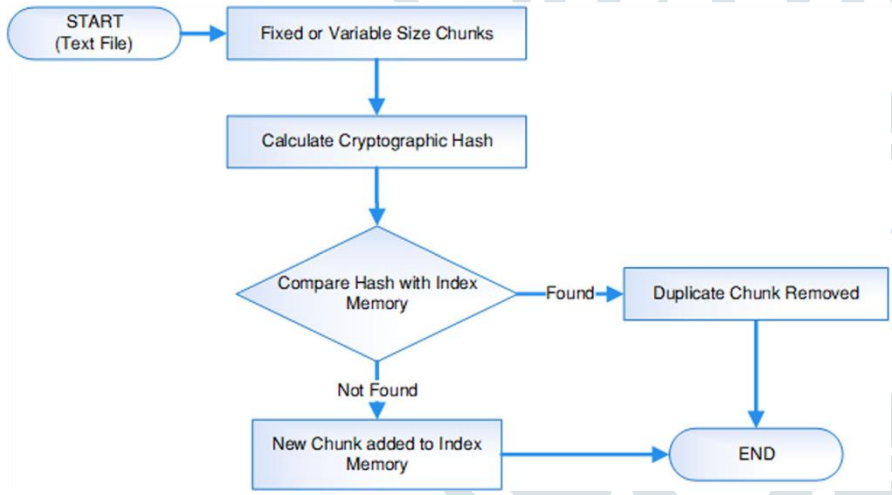

Fig. 2 Steps of deduplication process

An important role to find duplicate chunks In Fig. 2, the file processed for dedu-plication is first broken down into fixed- or variable-size blocks referred to as objects. Data deduplication compares and eliminates blocks that are of same fingerprints. The unique are stored and index is updated. The four generic steps of the deduplication process are as follows.

- A hash value is calculated first for each chunk of data using cryptographic hash function.
- A comparison is made between the hash values of chunks and existing hashes.
- The same hash values find duplicate chunk, and data are replaced with a logical pointer to the object already present in the database.

New chunk is added and index is updated.

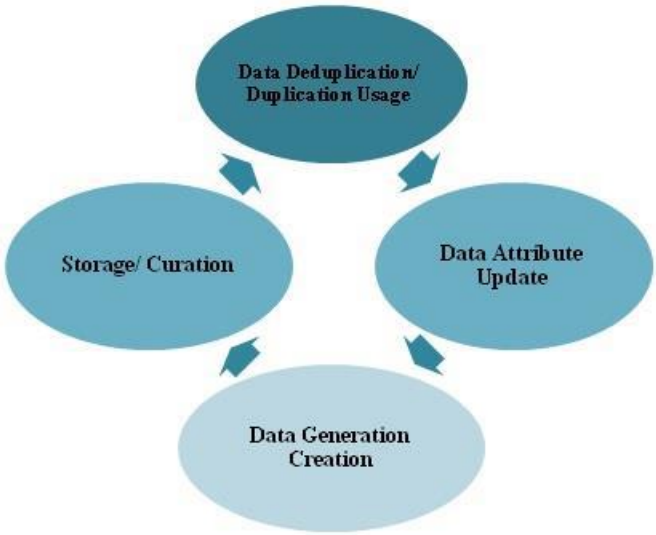

Fig.3. Life Cycle of Data Deduplication

Reliable re-use of this digital material, is only possible. If the digital duration, archiving and storage systems are well-defined and functioning with minimum resource to maximum returns. Hence, the control of these events in the Life Cycle is Deduplication process and securely of data are critical for any organization.

III. LITERATURE REVIEW

Initially the emphasis of researchers was on the reduction of data footprint by using disk rather than using tape for backups. From the start of year 2003 virtual tape libraries (VTL) were extensively used by the industry but the real advent of deduplication in fact, started in a 1970‟s when companies use to store large amount of customer contact information on tapes and had to eliminate duplicate entries. But, today deduplication is not only doing value addition in backup systems but additionally it is optimizing IOPs, SSD and DRAM efficiencies also. The various implementation of deduplication algorithm show that when a virtual machine disk is created which certain X amount of data already resting in it. It would not require to rewrite those X amount of data to the disk. Thus saving significant IOPs resources. Deduplication also improves the efficiency of SSD and DRAM as it able to maintain same piece of data (in case it is duplicate) in minimum space. This is significant saving as DRAM and SSD are more expensive than HDD. Then disaster recovery system also get more optimize with the aid of deduplication process. However, all these processes need highly efficient key management and optimized proof of ownership work flows for them to be successful. The following section gives a review on the latest strategies employed in this context for key management and [21] worked.

Junbeom Hur et al. [7] worked on the problem of building a secure key ownership schema that works dynamically with guaranteed data integrity against tag inconsistency attacks. The techniques used in this was the use of Re-encryption technique that enabled dynamic updates upon any ownership changes in the cloud storage [22]. Using this method, the authors claimed that the tag consistency becomes true and key management becomes more efficient in terms of computation cost as compare to RCE (Randomized convergent encryption). However the author did not focused their work on arithmetic validity of the keys. Although the lot of work has been done on ownership of keys. Chia-Mu Yu et al. [8] worked on the improvement of the cloud server and mobile device efficiency in terms of its storage capabilities and of key management scheme. The improvement was done using better flow of key management with bloom filter for managing memory without the need to access disk after storing data (Post Storage). The result of this paper claim reduced server user side latency. Jorge Blasco et al. [9] worked on improving the efficiency of resources (space, bandwidth, efficiency) and on improving the security during the deduplication process. The improvement was done using the working of bloom filter [23] and it application in key management scheme to thwart a malicious client attack that colluded with the legitimate owner of the files. The results of the paper show that resources were optimized and execution time was optimized when size of file grows. Thus the implemented algorithm provided a better tradeoff between space and bandwidth.

Jin Li et al. [10] worked on the problem of improving key management schema that it more optimal in eating resources and secure when key distribution operations occurs. The method employed to solve this challenge is use of the independent master key for encrypting the convergence keys and by avoiding its outsourcing. This is avoided by using RAMP secret sharing (RSSS) and dividing the duplication phases into small phases. The limitation of secret sharing is that each share of the secret must be least as large itself .The result is compression of the secret key is using hard due to randomness it process. But the random helps to keep the key secure and fresh also. Using this method the authors found a new key management scheme (Dekey) with help of ramp scheme that reduces the overhead (encoding and decoding) and is better than the previous scheme. Chao Yang et al. [11] the problem of the vulnerability of client side deduplication operation, this is when the Deduplication operations are done on the client machine or the place where the originally data resides. In case, the attacker pushes for access on unauthorized file stored on the server by just using file name and its hash value. The results of this paper show that scheme creates better provable ownership file operation that maintains high degree of detection power in terms of probability of finding unauthorized access to files.

In the research work (Xuexue Jin et al .[12]) the methods used information computed from shared file(s) .This way the data or convergent encryption remains vulnerable as it is based well known public algorithm. The techniques used by authors are deduplication encryption algorithms are combined with proof of ownership algorithm to achieve higher degree of security during

the deduplication process. The process also argument with proxy re-encryption (PRE) and digitalize credentials checks. The proxy re-encryption schemes basically allow the third party (called proxies) to modify the cipher which has been encrypted by one party so as to it may decipher. Using these methods the authors achieved anonymous deduplication encryption along with key management test, consequently the level of protection was increased and attacks were avoided. Danny Harnik et al. [13] worked on improving the cross user (s) interaction securely with higher degree of privacy during deduplication. The improvement was done using the multiple methods that included: - a). Stop cross over user interaction. b). Allow user to use their own private keys to encrypt. c). Randomized algorithm. The result in the paper shows a reduced cost of operations to secure the duplication process. Reduced leakage of information during deduplication process. Higher degree of fortification. Jingwei Li et al. [14] worked the problem of integrity auditing and security of deduplication and The authors have proposed and implemented two methods viz Sec Cloud and Sec Cloud+, both systems improve auditing the maintain ace with help of map reduce architecture. The results show that the implementation provided better performance of periodic integrity check and verification without the local copy of data files. The process also provides better degree of proof of ownership process integrated with auditing. Kun He et al. [15] in this research the problem of reducing complications due to structure diversity and private tag generation and find better alternatives to homomorphic authenticated tree. (HAT). Since, the homomorphic authenticated tree are allows complex computation without compromising the security level .The term is implied for same structure data keys here for securing the Deduplication process. This encryption is play an important part in cloud computing that gives change to companies to store encrypted data .The method used to solve the problem is the use random oracle model to avoid occurrence of breach and constructs to do unlimited number of verifications and update operations. DeyPoS which means deduplicable dynamic proof of storage. Using this method the authors claimed that the theoretical and experimental results show that the algorithm (DeyPoS) implementation is highly efficient in conditions where the file size grows exponentially and large number of blocks are there. Jin Li et al. [16] the researchers are worked to provide a better protected data, and reduce duplication copies in storage with help of encryption scheme and find alternate deduplication method. The techniques used by them to solve this problem with use of hybrid cloud [24] architecture for higher degree of security (token based). The token were used to maintain storage that does not have deduplication and it is more secure due to its dynamic behavior. The results claimed in the paper shows that the implemented algorithm gives minimal overhead compared to the normal operations. Zheng Yan et al. [25] the researcher had put effort on reducing the complexity of key management steps during data duplication process. The improvement was done using with less complex encryption with better level of security. This is done with the help of Attribute Based Encryption algorithm (ABE). The results shown in this paper claim to reduce complexity overhead and execution time when file size grows as compared to preview work.

IV. CHALLENGES FOUND

The following section gives important pointers that need further attention, investigation and analysis. After reading prominent works in this context, it can be inferred that the degree of issues related to the implementation of crypto algorithm [26] in terms of their mathematical function, is not that difficult is embracing and applying these cryptographic mathematics to real time current technological changes takes place in storage technology. These days machine to machine communication is increasing many folds and man to man communication in collaborative environment is also increasing at exponential rate. In fact multiple teams from multiple time zones in collaborative to, generate lot of data that may be work duplicate in nature. Protecting such data and for maintaining its integrity remains a challenge due to cyber criminals. Hence security policy key generation schemes, proof of ownership scheme need careful security arrangements. Specially, in cases, where distributed storage is a norm. Such scenario also necessitates the need of centralized anonymous credentials validity at each state of deduplication [27] work flow.

In certain scenario there is also a need to eliminate a trusted credentials issuer so as to stream line the deduplication work flow in checking credential without compromising security level. Deduplication process may also require further simplification of method involved in validity of keys. From literature survey, it was also found that most of solution in deduplication work at block level, but in certain cases zone level deduplication may be better for storage environments. However, empirical studies, implementation of such strategy remains slow. May be this is due to security concerns, but it is apparent that zone level deduplication will require less number of validity checks leading to lower overhead. This is due to the fact that when the block size is small, a large hash table needs to be maintain which makes hash table ultimately unwieldy. Another debate that can be found in current literature is about whether to use semantic deduplication or global deduplication. Ultimately deduplication ratios and overhead needs to be optimized and it has been found that semantic deduplication process require high grade hardware resources. When a global approach based is used to remove redundant file again a large overhead may be involved. Hence it remains a challenge to decide at what level the deduplication process should be implemented, especially in cases where key management and schemes become complex having multiple functional calls leading to high overhead. Next challenge worth mentioning here is about choosing appropriate type of deduplication process. Empirical studies have shown that source deduplication solution works well, when servers have adequate resources allocated for deduplication process and key management [28] mechanism. But, normally host machines rarely can allocate large amount of resources to deduplication process. Hence, sometimes a target device based deduplication may be a better alternative. But this strategy is not without a challenges because the target machine may become over whelmed and it may not be able to keep pace with the demand of deduplication threads issued from server request. Hence, finding right or appropriate deduplication methods particular scenario requires careful planning and understanding of resources and algorithms involved. Many methods compute the secret key based on Recursive method, which have more overhead as compared to the methods that are vectorized. Some of the vectorized implementations of such algorithms can be improved by reducing the number of steps with one line computational methods, especially when the powers of exponent are smaller than 8. Many algorithms for exponentiation do not provide defense against side-channel attacks, when deduplication process is run over

network. An attacker observing the sequence of squaring and multiplications can (partially) recover the exponent involved in the computation.

V. CONCLUSION

In this paper, sections have been dedicated to the discussion on the values, concepts that need to be understood to overcome the challenges in deduplication algorithms implementations. It was found that at each level of duplication process (file, block, chunk, zone) there is a needs of keys to be arithmetically valid and there ownership also need to be proved for proper working of any secure (Source, Target, Semantic ,Local, Hardware etc.) Deduplication system. The process becomes prone to attacks, when the process is applied in geo-distributed storage architecture. The complexity for cheating ownership verification is as difficult as performing strong collision attack of the hash function due to these mathematical functions. Finding the discrete algorithm of a random elliptic curve element with respect to a publicly known base point is infeasible this is (ECDLP). The security of the elliptic curve cryptography depends on the ability to compute a point multiplication and the mobility to compute the multiple given the original and product points. The size of the elliptic curve determines the difficulty of the problem.
From this study, it can also concluded that there is no absolute or perfect solution of deduplication. One of the main criteria are security and resources required. Major developments in cloud, Storage can only move forward if optimization of disk space is done The Global approach of Deduplication is slow but may overall remove number of duplicate files. Local strategy may be fast but give less number removals. In certain cases, there is need to exploit the use of GPUs for speedy operations in both the cases. And all these types of disk spaces need to be secure with help of Key management. Both these will be dysfunctional without an operation to check the validity of the keys. Modular arithmetic is normally used to create groups of keys, rings and fields which are fundamental building blocks of most modern public-key cryptosystems. The reason is that modular arithmetic gives a chance to increase difficulty in guessing the keys if we introduce modular reduction for example in Key management.

VI. FUTURE SCOPE

As discussed, in the above section mathematical methods may be used for doing computations related to arithmetic validity of the keys produced for security purpose as it involves easier steps and reduce the number of bits required for doing multiplication operations etc. Other than this, the future research work to apply security network need of sensors that have low memory and computational power to run expensive cryptography operations such as public key validation and key exchange thereafter.

REFERENCES

[1] K. Zarour and N. Zarour, "Data center strategy to increase medical information sharing in hospital information," International Journal of Information Engineering and Electronics Busniness, vol. 5, p. 33, 2013.
[2] M. Portolani, M. Arregoces, D. W. Chang, N. A. Bagepalli and S. Testa, "System for SSL re-encryption after load balance," 2010.
[3] D. Harnik, B. Pinkas and A. Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud storage," IEEE Security \& Privacy, vol. 8, pp. 40--47, 2010.
[4] J. Li, J. Li, D. Xie and Z. Cai, "Secure auditing and deduplicating data in cloud," IEEE Transactions on Computers, vol. 65, pp. 2386--2396, 2016.
[5] F. shieh , M. G. Arani and M. Shamsi, "An extended approach for efficient data storage in cloud computing environment," International Journal of Computer Network and Information Security, vol. 7, p. 30, 2015.
[6] S. Gupta, A. Goyal and B. Bhushan, "Information hiding using least significant bit steganography and cryptography," International Journal of Modern Education and Computer Science, vol. 4, p. 27, 2012.
[7] K. V. K. and A. R. K. P. , "Taxonomy of SSL/TLS Attacks," International Journal of Computer Network and Information Security, vol. 8, p. 15, 2016.
[8] J. M. Sundet, D. G. Barlaug and T. M. Torjussen, "The end of the Flynn effect?: A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century," Intelligence, vol. 32, pp. 349--362, 2004.
[9] W. Lawrence and S. Sankaranarayanan, "Application of Biometric security in agent based hotel booking system-android environment," International Journal of Information Engineering and Electronic Business, vol. 4, p. 64, 2012.

[10] N. Asokan, V. Niemi and P. Laitinen, "On the usefulness of proof-of-possession," in Proceedings of the 2nd Annual PKI Research Workshop, 2003, pp. 122--127.
[11] X. Jin, L. Wei, M. Yu, N. Yu and J. Sun, "Anonymous deduplication of encrypted data with proof of ownership in cloud storage," in Communications in China (ICCC), 2013 IEEE/CIC International Conference on, 2013, pp. 224--229.
[12] Z. Yan, M. Wang, Y. Li and A. V. Vasilakos, "Encrypted data management with deduplication in cloud computing," IEEE Cloud Computing, vol. 3, pp. 28--35, 2016.
[13] D. Whitfield and M. E. Hellman, "New directions in cryptography," IEEE transactions on Information Theory, vol. 22, pp. 644--654, 1976.
[14] H. Riesel, "Prime numbers and computer methods for factorization," vol. 126, Springer Science \& Business Media, 2012.
[15] R. A. Patel, M. Benaissa, N. Powell and S. Boussakta, "Novel power-delay-area-efficient approach to generic modular addition," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 54, pp. 1279--1292, 2007.

[16] "Repeated Squaring," Wednesday March 2017. [Online]. Available: http://www.algorithmist.com/index.php/Repeated_Squaring. [Accessed Wednesday March 2017].
[17] "Calculating Powers Near a Base Number," Wednesday March 2017. [Online]. Available: http://www.vedicmaths.com/18-calculating-powers-near-a-base-number. [Accessed Wednesday March 2017].
[18] C.-M. Yu, C.-Y. Chen and H.-C. Chao, "Proof of ownership in deduplicated cloud storage with mobile device efficiency," IEEE Network, vol. 29, pp. 51--55, 2015.
[19] J. Hur, D. Koo, Y. Shin and K. Kang, "Secure data deduplication with dynamic ownership management in cloud storage," IEEE Transactions on Knowledge and Data Engineering, vol. 28, pp. 3113--3125, 2016.
[20] J. Blasco, R. D. Pietro, A. Orfila and A. Sorniotti, "A tunable proof of ownership scheme for deduplication using bloom filters," in Communications and Network Security (CNS), 2014 IEEE Conference on, 2014, pp. 481--489.
[21] J. L. a. Y. K. L. a. X. C. a. P. P. C. L. a. W. Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication," IEEE Transactions on Parallel and Distributed Systems, vol. 26, pp. 1206-1216, 2015.
[22] K. H. a. J. C. a. R. D. a. Q. W. a. G. X. a. X. Zhang, "DeyPoS: Deduplicatable Dynamic Proof of Storage for Multi-User Environments," IEEE Transactions on Computers, vol. 65, pp. 3631-3645, 2016.
[23] A. Kumar and A. Kumar, "A palmprint-based cryptosystem using double encryption," in SPIE Defense and Security Symposium, 2008, pp. 69440D--69440D.
[24] C. Yang, J. Ren and J. Ma, "Provable ownership of files in deduplication cloud storage," Security and Communication Networks, vol. 8, pp. 2457--2468, 2015.
[25] J. Li, X. Chen, M. Li, J. Li, P. P. Lee and W. Lou, "Secure deduplication with efficient and reliable convergent key management," IEEE transactions on parallel and distributed systems, vol. 25, pp. 1615--1625, 2014.
[26] S. P. Dwivedi, "An efficient multiplication algorithm using Nikhilam method," 2013.
[27] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," IEEE Transactions on information theory, vol. 23, pp. 337--343, 1977.
[28] W. Xia, H. Jiang, D. Feng, F. Douglis, P. Shialane, Y. Hua, M. Fu, Y. Zhang and Y. Zhou, "A comprehensive study of the past, present, and future of data deduplication," Proceedings of the IEEE, vol. 104, pp. 1681--1710, 2016.

[29] P. KUMAR, M.-L. LIU, R. VIJAYSHANKAR AND P. MARTIN, "SYSTEMS, METHODS, AND COMPUTER PROGRAM PRODUCTS FOR SUPPORTING MULTIPLE CONTACTLESS APPLICATIONS USING DIFFERENT SECURITY KEYS," 2011.